

# **EXHIBIT B**

12/6/2024

Richard Kadrey, et al. v. Meta Platforms, Inc. Nikolay Bashlykov 30(b)(6)  
Highly Confidential - Attorneys' Eyes Only

Page 1

UNITED STATES DISTRICT COURT  
NORTHERN DISTRICT OF CALIFORNIA  
SAN FRANCISCO DIVISION

---

RICHARD KADREY, et al.,	)	
Individual and	)	
Representative	)	
Plaintiffs,	)	
	)	
v.	)	Case No.:
	)	3:23-cv-03417-VC
META PLATFORMS, INC.,	)	
Defendant.	)	
	)	

---

\*\* HIGHLY CONFIDENTIAL - ATTORNEYS' EYES ONLY \*\*

Videotaped 30(b)(6) deposition of Defendant  
META PLATFORMS, INC.,  
by and through its corporate designee  
NIKOLAY BASHLYKOV  
Friday, December 6, 2024

London, England  
United Kingdom

Reported stenographically by:  
Leah M. Willersdorf,  
RMR, CRR, FBIVR, ACR, QRR2\*, CLR

---

DIGITAL EVIDENCE GROUP  
1730 M. Street, NW, Suite 812  
Washington, D.C. 20036  
(202) 232-0646

12/6/2024

Richard Kadrey, et al. v. Meta Platforms, Inc.  
Highly Confidential - Attorneys' Eyes Only

Nikolay Bashlykov 30(b)(6)

Page 66

1	Q. I believe these datasets were produced as	13:58:07
2	training data for LLaMAs 1 through 3, and it's your	13:58:11
3	testimony that you can't confirm whether that's true?	13:58:21
4	MR. WEINSTEIN: Object to form.	13:58:26
5	THE WITNESS: What I was saying is that	13:58:33
6	for some, like, there is a term which is called	13:58:34
7	"poking" --	13:58:44
8	BY MS. POUEYMIROU:	13:58:44
9	Q. Mmm-hmm.	13:58:44
10	A. -- the data, so I wouldn't exclude that	13:58:46
11	for some of the dataset the poking was less than 1,	13:58:48
12	that is possible.	13:58:52
13	Q. Okay. Was B3G -- what is B3G? You	13:58:53
14	test -- what is B3G?	13:59:00
15	A. In context of these files?	13:59:02
16	Q. Mmm-hmm.	13:59:04
17	A. It's Books3 Gutenberg.	13:59:05
18	Q. And how has Books3 Gutenberg been used for	13:59:09
19	training Meta LLaMA models?	13:59:15
20	A. When you're asking how, what specifically	13:59:20
21	are you asking about?	13:59:23
22	Q. Has Books3 Gutenberg been used for	13:59:25

12/6/2024

Richard Kadrey, et al. v. Meta Platforms, Inc.  
Highly Confidential - Attorneys' Eyes Only

Nikolay Bashlykov 30(b)(6)

Page 67

1	pretraining Meta LLaMA models?	13:59:30
2	A. It was.	13:59:31
3	Q. Has Books3 Gutenberg been used in	13:59:32
4	post-training of Meta LLaMA models?	13:59:36
5	A. Parts of it was, yes.	13:59:38
6	Q. Do you know which parts?	13:59:41
7	A. It's hard to say, and you can't say	13:59:49
8	document by document, so subset of Books3 was used.	13:59:57
9	Q. Was Books3 Gutenberg, B3G, used for	14:00:01
10	ablation studies?	14:00:07
11	A. For some of the ablation studies, parts of	14:00:11
12	Books3G documents were used.	14:00:19
13	Q. And were those ablation studies in the	14:00:20
14	pretraining part of the life cycle?	14:00:23
15	A. Yes.	14:00:35
16	Q. And were they the ablation studies in the	14:00:35
17	post-training part of the life cycle?	14:00:39
18	A. Yes.	14:00:41
19	Q. And was B3G used for benchmarking? Is	14:00:42
20	that the same as ablations?	14:00:48
21	A. Could you clarify on the benchmarking what	14:00:56
22	do you mean?	14:00:58

12/6/2024

Richard Kadrey, et al. v. Meta Platforms, Inc.  
Highly Confidential - Attorneys' Eyes Only

Nikolay Bashlykov 30(b)(6)

Page 68

1 Q. How do you define "benchmarking"? It's a 14:00:59

2 term in the documents frequently. 14:01:03

3 A. Mmm. So it depends on the context, but 14:01:08

4 most -- I would refer it as measuring performance of 14:01:11

5 the model against certain benchmarks. 14:01:16

6 Q. Is that like an ablation study -- 14:01:20

7 A. So ablation study, part of the ablation 14:01:24

8 study is measuring the performance on certain 14:01:31

9 benchmarks, like, for specific ablations. 14:01:34

10 Q. Okay. So ablation studies and 14:01:37

11 benchmarking are not always synonymous? 14:01:45

12 A. They are not. 14:01:45

13 Q. Okay. 14:01:46

14 Was B3G used to test memorization? 14:01:52

15 MR. WEINSTEIN: Object to form, outside 14:01:56

16 the scope. 14:01:57

17 MS. POUEYMIROU: No. Topic 7 that you've 14:01:57

18 agreed to is the uses of these datasets. 14:02:00

19 MR. WEINSTEIN: In training the LLaMA 14:02:02

20 models. 14:02:03

21 MS. POUEYMIROU: Memorization is a part of 14:02:05

22 training. 14:02:07

12/6/2024

Richard Kadrey, et al. v. Meta Platforms, Inc.  
Highly Confidential - Attorneys' Eyes Only

Nikolay Bashlykov 30(b)(6)

Page 69

1	MR. WEINSTEIN: I'm not sure I agree with	14:02:07
2	that. Again, not for here. I'll put the objection on	14:02:08
3	the record. The witness can answer to the best of his	14:02:12
4	knowledge.	14:02:15
5	BY MS. POUEYMIROU:	14:02:15
6	Q. Let me actually ask: Is -- well, we'll	14:02:16
7	stick with that question. Has B3G been used for	14:02:23
8	testing memorization?	14:02:27
9	MR. WEINSTEIN: Same objection; outside	14:02:29
10	the scope.	14:02:40
11	THE WITNESS: To the best of my knowledge,	14:02:40
12	some parts or documents of B3G could be used in	14:02:42
13	memorization, like the project which checks.	14:02:58
14	I wouldn't call that memorization and, probably,	14:03:02
15	I don't know, like, the full term that you -- that	14:03:04
16	copies it, but...	14:03:12
17	BY MS. POUEYMIROU:	14:03:12
18	Q. Like the project which checks for	14:03:12
19	memorization; is that what you meant?	14:03:14
20	A. So the --	14:03:16
21	MR. WEINSTEIN: Same objection; outside	14:03:17
22	the scope.	14:03:19

12/6/2024

Richard Kadrey, et al. v. Meta Platforms, Inc.  
Highly Confidential - Attorneys' Eyes Only

Nikolay Bashlykov 30(b)(6)

Page 70

1	You can answer.	14:03:20
2	THE WITNESS: What I'm referring to	14:03:23
3	memorization is the project that was checking if,	14:03:25
4	given a small subsequence of the document, how does	14:03:30
5	the generation of a particular model match or does not	14:03:34
6	match the preceding subsequence, usually limited to	14:03:42
7	50 tokens or so.	14:03:48
8	BY MS. POUEYMIROU:	14:03:49
9	Q. And to prevent memorization in the model,	14:03:49
10	does that involve training the model?	14:03:52
11	MR. WEINSTEIN: Object to form.	14:04:01
12	THE WITNESS: I don't really understand	14:04:03
13	the question.	14:04:07
14	BY MS. POUEYMIROU:	14:04:07
15	Q. My question is that in order to prevent	14:04:08
16	memorization, does that involve training? Is the	14:04:11
17	model trained to not regurgitate memorized text, or	14:04:22
18	does Meta achieve that in a different way?	14:04:27
19	MR. WEINSTEIN: Object to form, compound.	14:04:30
20	THE WITNESS: I would say that memorize --	14:04:42
21	like, the project that I'm referring to which measures	14:04:47
22	a specific metric of how many tokens does the model	14:04:49

12/6/2024

Richard Kadrey, et al. v. Meta Platforms, Inc.  
Highly Confidential - Attorneys' Eyes Only

Nikolay Bashlykov 30(b)(6)

Page 71

1	generate, having the prompt of specific tokens from	14:04:53
2	the document, so that particular project, to prevent,	14:04:58
3	like, or lower or somehow control the number of tokens	14:05:12
4	which the model generates or, like, the matching kind	14:05:19
5	of ratio --	14:05:22
6	BY MS. POUEYMIROU:	14:05:22
7	Q. Mmm-hmm.	14:05:22
8	A. -- usually is controlled by	14:05:25
9	limiting/epoching of certain datasets, so it's kind of	14:05:31
10	like it's an opposite of training, it's basically	14:05:35
11	reducing the amount of times the document is seen by	14:05:40
12	the model during training.	14:05:43
13	Q. Okay. Has LibGen been used by Meta for	14:05:44
14	pretraining in relation to any of its Meta LLaMA	14:05:52
15	models?	14:05:57
16	A. Parts of LibGen were used in pretraining	14:06:01
17	of one of the LLaMA models.	14:06:06
18	Q. Was sci-mag used?	14:06:09
19	A. Parts of sci-mag were used.	14:06:16
20	Q. Okay. Has LibGen been used in the	14:06:19
21	post-training of any of the Meta LLaMA models?	14:06:26
22	A. Parts of LibGen dataset were used as a	14:06:37

12/6/2024

Richard Kadrey, et al. v. Meta Platforms, Inc.  
Highly Confidential - Attorneys' Eyes Only

Nikolay Bashlykov 30(b)(6)

Page 72

1	kind of source files to prepare synthetic datasets	14:06:42
2	in -- for fine tuning.	14:06:54
3	Q. Okay. Is that MMLibU?	14:06:57
4	A. Sorry, what is MMLibU?	14:07:03
5	Q. Well, we'll look at a document.	14:07:06
6	Has LibGen been used -- and when you say	14:07:08
7	"parts," how would we determine which parts? What	14:07:11
8	would we need to look at?	14:07:14
9	A. So when I was saying "parts," is that not	14:07:21
10	all of the documents in the dataset were used for	14:07:25
11	fine tuning because fine tuning usually requires much	14:07:36
12	less tokens for training.	14:07:40
13	Q. And so where would you -- would Meta then	14:07:42
14	create a smaller subset of LibGen for the purpose of	14:07:48
15	fine tuning?	14:08:02
16	A. So for the purpose of fine tuning, you	14:08:02
17	need to adjust the format of the documents that the	14:08:05
18	model is trained on; so, for that purpose, if the	14:08:09
19	source is LibGen, then -- or parts of LibGen, then	14:08:14
20	these documents would be processed to produce a	14:08:23
21	smaller subset of LibGen --	14:08:29
22	Q. And would there be source code pertaining	14:08:33

12/6/2024

Richard Kadrey, et al. v. Meta Platforms, Inc.  
Highly Confidential - Attorneys' Eyes Only

Nikolay Bashlykov 30(b)(6)

Page 73

1 to creating that smaller set? 14:08:36

2 A. There should be in the code base. 14:08:40

3 Q. And where would that be stored? In which 14:08:45

4 source code repository would that be stored in? 14:08:48

5 A. I think -- sorry. Most likely, it 14:08:57

6 should be in XLFormers -- 14:09:01

7 Q. Okay. 14:09:06

8 A. -- but I can, yeah, check if that is 14:09:07

9 needed. 14:09:10

10 Q. Okay. Has LibGen been used for ablation 14:09:10

11 studies? 14:09:21

12 MR. WEINSTEIN: Object to form. 14:09:21

13 BY MS. POUEYMIROU: 14:09:21

14 Q. You can answer. 14:09:22

15 A. Could you specify which ablation studies 14:09:23

16 or, like, in general? 14:09:26

17 Q. For any ablation studies. Have you used 14:09:27

18 LibGen for any ablation studies? 14:09:30

19 MR. WEINSTEIN: Object to form, scope. 14:09:33

20 You can answer. 14:09:35

21 THE WITNESS: For some of the ablation 14:09:38

22 studies, LibGen was used, yes. 14:09:41

12/6/2024

Richard Kadrey, et al. v. Meta Platforms, Inc.  
Highly Confidential - Attorneys' Eyes Only

Nikolay Bashlykov 30(b)(6)

Page 74

1 BY MS. POUEYMIROU: 14:09:43

2 Q. And were those ablation studies in the 14:09:43

3 pretraining part of a life cycle? 14:09:45

4 A. Yes. 14:10:00

5 Q. And were there ablation studies using 14:10:01

6 LibGen in the post-training part of the life cycle? 14:10:04

7 A. Yes. 14:10:08

8 Q. Okay. And has LibGen been used for 14:10:08

9 benchmarking? 14:10:11

10 A. So, again, if you could specify what do 14:10:12

11 you mean under this? 14:10:15

12 Q. Okay. We'll look at a document. 14:10:16

13 Has LibGen been used to determine whether 14:10:19

14 to license with different content providers? 14:10:21

15 MR. WEINSTEIN: Object to form, outside 14:10:25

16 the scope. 14:10:28

17 MS. POUEYMIROU: It's a use of LibGen. 14:10:28

18 MR. WEINSTEIN: I -- hard to believe you 14:10:31

19 could have it with a straight face. 14:10:34

20 You can answer that question if you can. 14:10:36

21 Object to form, outside the scope. 14:10:38

22 THE WITNESS: I do not know how LibGen was 14:10:47